Assessing Fair Policing in Austin, TX

Team FunkyStats

4/18/2021

Introduction

This study investigates racial disparities in traffic stops by the Austin Police Department. Using data available from Austin Open Data, a Texas government-run data portal, and from the Stanford Open Policing Project, we evaluate these disparities using models derived from the "hit rate" and the effect of the "veil of darkness," two often-cited methods for assessing fair policing.

Introduction

Our study consists of three parts:

- Exploratory data analysis to get a big picture of policing in Austin:
  - Benchmark Test
  - Outcome Test
  - Veil of Darkness Test
- Various modeling strategies to assess the severity of racial disparities:
  - Logistic Regression
  - Bayesian Hierarchical Model
- Propose a measure of fairness
  - based on the differences in the posterior median hit rate among individual police officers

Available Data

- Stanford Open Policing Project data (2006.01.01 - 2016.06.30, 463,944 stops): stops time, the driver race, searched or frisked, contraband discovered etc.
    - Merits: contain driver race
    - Drawbacks: miss time and location information
- APD Racial Profiling data (2019, 79,693 stops): similar to Stanford data
    - Merits: contain time, location, and officer race
    - Drawback: miss driver race
- US census demographic data (2012-2017 5-year average data, 2019): contain population of different races
- APD Racial Profiling Report: contain driver races in general sense

Summary Statistics

	nobs	nmis	uniq	mean	SD	min	25%	50%	75%	max
subject_age	480091	3164	94	37.98	13.82	10.00	26.00	36.00	48.00	103.00
subject_sex	482881	374	2	0.30	0.46	0.00	0.00	0.00	1.00	1.00
frisk_performed	483255	0	2	0.02	0.15	0.00	0.00	0.00	0.00	1.00
search_conducted	483255	0	2	0.04	0.20	0.00	0.00	0.00	0.00	1.00
search_person	483255	0	2	0.03	0.18	0.00	0.00	0.00	0.00	1.00
search_vehicle	483255	0	2	0.02	0.15	0.00	0.00	0.00	0.00	1.00

	nobs	nmis	uniq	mean	SD	min	25%	50%	75%	max
contraband_found	19256	0	2	0.25	0.43	0.00	0.00	0.00	0.00	1.00
contraband_drugs	19256	0	2	0.01	0.12	0.00	0.00	0.00	0.00	1.00
contraband_weapons	19256	0	2	0.05	0.21	0.00	0.00	0.00	0.00	1.00
frisk_performed	19256	0	2	0.51	0.50	0.00	0.00	1.00	1.00	1.00

- Summary statistics for all stops
- Summary statistics for for stops during which a search was performed

Summary Statistics

Table 1: Subject race.

Race	n	percent
asian/pacific islander	13167	0.0272466
black	72324	0.1496607
hispanic	123943	0.2564764
other	2626	0.0054340
unknown	3135	0.0064873
white	268058	0.5546950
NA	2	NA

Summary Statistics

Table 2: Search basis.

Search basis	n	percent
consent	3195	0.1659223
other	276	0.0143332
plain view	152	0.0078936
probable cause	15633	0.8118509
NA	463999	NA

Summary Statistics

We note that the distribution of stops per officer has an extremely long tail.

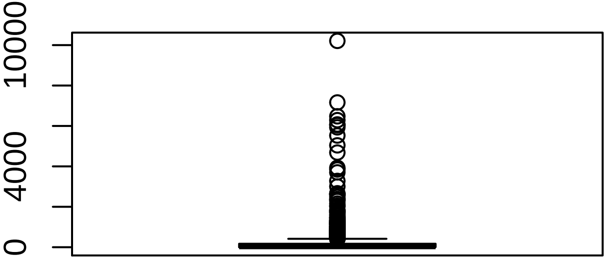


Figure 1: Distribution of stops by unique officer ID

Exploratory Analysis

Examine the count of stops by race during 2006-2015:

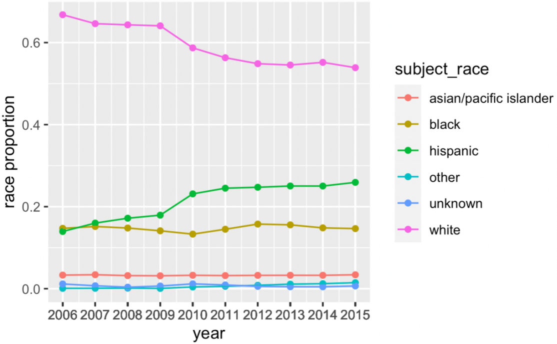
- Half of the stops involved were of white subjects, about four times the number of stops of black people
- The white population in Austin (445,269) is almost 7 times than the black population (66,724)

Driver Race	Counts	Proportion
asian/pacific	11658	0.033
black	52381	0.147
hispanic	765707	0.215
other	2105	0.006
unknown	2622	0.007
white	211588	0.593

Exploratory Analysis

Examine race proportion in each year:

- Annual trends are very different by race
- Fewer white drivers stopped especially after 2009
- An increasing trend of Hispanic and black drivers being stopped



Benchmark Test

$$\text{Stop Rate}_i = \frac{\text{Number of Stops for Race } i}{\text{Population of Race } i}$$

$$\text{Search Rate}_i = \frac{\text{Number of Stopped People Who Were Searched for Race } i}{\text{Number of Stops for Race } i}$$

$$\text{Frisk Rate}_i = \frac{\text{Number of Stopped People Who Were Frisked for Race } i}{\text{Number of Stops for Race } i}$$

Driver Race	Counts	Population	Proportion	Stop Rate	Search Rate	Frisk Rate	Hit Rate
asian/pacific	11658	63752	0.033	0.183	0.015	0.011	0.188
black	52381	66724	0.147	0.785	0.092	0.039	0.254
hispanic	765707	316709	0.215	0.242	0.086	0.044	0.323
white	211588	445269	0.593	0.475	0.031	0.021	0.318

Benchmark Test Caveats

- The racial disparity by the police is clear from benchmark test, but it is insufficient evidence of discriminative policing.
    - E.g., if black drivers, hypothetically, spend more time on the road than white drivers, that could explain the higher stop rates for black drivers
- The key part of this analysis is to find out the true distribution of the drivers violating the traffic laws or conducting crimes.
- We need to check if different race groups are disproportionately stopped corresponding to their rates of violating the law.

Outcome Test

- Define a successful search as one that uncovers contraband
- Hit rate is the proportion of searches that are successful
  - If racial groups have different hit rates, it can be taken as evidence of discriminative policing

$$\text{Hit Rate}_i = \frac{\text{Number of Contraband Uncovered for Race } i}{\text{Number of Searched People for Race } i}$$

Driver Race	Counts	Population	Proportion	Stop Rate	Search Rate	Frisk Rate	Hit Rate
asian/pacific	11658	63752	0.033	0.183	0.015	0.011	0.188
black	52381	66724	0.147	0.785	0.092	0.039	0.254
hispanic	765707	316709	0.215	0.242	0.086	0.044	0.323
white	211588	445269	0.593	0.475	0.031	0.021	0.318

Outcome Test Caveats

- Only outcomes available: Although the outcome test is simple and intuitive, the actual threshold for searching someone is not observed.
- Infra-marginality problem: Outcome tests of disparate treatment may only measure the average outcome and not the outcomes associated with the marginal decision. Observing that the average hit rate for minorities was lower than for whites does not necessarily prove that the threshold (or marginal) expected success rate was lower for minorities than for whites
- Subgroup validity problem: When a particular observable characteristic is valid for some races but not for others, it is possible that a decisionmaker who conditions decisions on this characteristic generally might induce racially disparate outcomes. A decisionmaker's unwillingness to engage in disparate racial treatment may induce the racial disparities in outcomes

Veil of Darkness Test

- Hypothesis: officers who are engaged in racial profiling are less likely to be able to identify a driver's race after dark than during daylight
- Under this hypothesis, if stops made after dark had smaller proportion of black drivers stopped than stops made during daylight, it could be evidence of racial profiling.
- Two key elements: Driver race & Stop time
- Alternative: measure the racial population in different areas through zip codes. If the number of the stops made during daytime and nighttime in black populated ares is significantly different from the ones made in white populated area, it could be evidence of racial profiling.

Veil of Darkness Test

In order to accurately distinguish the daytime and nighttime, we compute the daily subset and dusk time for Austin in 2019.

- Earliest sunset in 2019 was at around 17:32 in early December and it goes fully dark in 26 minutes
- Latest sunset time was around 20:38 late June and it was fully dark after 28 minutes.

Date	Sunset	Dusk	Sunset Minute	Dusk Minute
2019-12-02	17:31:42	17:57:48	1051	1077
2019-12-01	17:31:45	17:57:48	1051	1077
2019-06-30	20:37:58	21:05:27	1237	1265
2019-06-29	20:37:56	21:05:27	1237	1265

- Daytime Stop: stops happening before sunset
- Nighttime Stop: stop happening after the dusk

Veil of Darkness Test

According to ZIP codes and the corresponding demographic data, we consider

- Black Dominant Area (BDA): the areas consist of more black people
- White Dominated Area (WDA): the areas consist of more white people

For simplicity of the analysis, here we consider only the black and the white population groups. Hence, each zip code is regarded as a location with label as white (WDA) or black (BDA).

Veil of Darkness Test

	Day	Night
BDA	124	126
WDA	2937	2216

- Assume two rows as independent binomial samples
- Of $n_1 = 250$ recorded stops in black dominated area, 124 stops happened during the daytime, a proportion of $p_1 = 124/250 = 0.496$
- Of $n_2 = 5153$ recorded stops in white dominated area, 124 stops happened during the daytime, a proportion of $p_2 = 2937/5153 = 0.570$
- The sample difference of proportions is 0.074
- We obtain Fisher's exact test for testing null hypothesis of independence of the two rows with p value of 0.02, indicating the strong evidence that the police are not equally likely practicing during day and night to different racial groups.

Veil of Darkness Test Caveats

- Artificial lighting (e.g., from street lamps) can weaken the relationship between sunlight and visibility, and so the method may underestimate the extent to which stops are predicated on perceived race.
- Vehicle make, year, and model often correlate with race and are still visible at night, which could lead to the test under-estimating the extent of racial profiling.
- The test doesn't control for stop reason, which is often correlated with both race and time of day.
    - E.g: broken tail light stops

Hit Rate and Causal Issues

- Unmeasured confounders: Crime rates are known to be correlated with income and demographic factors
  - More officers patrolled areas with higher crime rates
  - Neighborhoods with higher minority populations is expected to see more minority traffic stops
- Although this still exposes problems in Austin, it could be interpreted as a problem of economic segregation, not traffic fairness.
- To overcome this problem, we propose to look at the hit rate with more details in the modeling parts:
  - The probability of finding contraband items should be equal among all races, regardless of the neighborhood that the search conducted
  - Using hit rate does not eliminate all unmeasured confounders, but it helps mitigate the problem

Logistic Regression for Frisk Rate

Our descriptive analysis shows that black people in Austin seem to be more likely to be stopped by the police. We want to answer the question, given a person is stopped, what factors may impact the likelihood of that person being frisked? To investigate this, we fit a logistic regression model with `frisk` as the dependent variable and `race`, `age`, and `sex`.

$$\text{Logit}[\text{P(Being Frisked)}] = \beta_0 + \beta_1\text{Race} + \beta_2\text{Age} + \beta_3\text{Sex}$$

Logistic model for frisk rate vs. race, age, and sex

term	estimate	std.error	statistic	p.value
(Intercept)	-2.984	0.102	-29.402	0.000
subject_raceblack	1.503	0.100	15.044	0.000
subject_racehispanic	1.310	0.099	13.214	0.000
subject_racewhite	0.719	0.099	7.260	0.000
subject_raceother	0.617	0.180	3.434	0.001
subject_raceunknown	0.647	0.183	3.544	0.000
subject_age	-0.046	0.001	-51.125	0.000
subject_sexfemale	-1.643	0.036	-45.311	0.000

Logistic Regression for Contraband found

We want to investigate how likely contraband items are found when searching is performed. This is equivalent to calculating hit rate defined in section 2.2.2. We argue that if racial bias does not exist, the hit rate should be equal for all races. In other words, we expect to find that `race` is not an essential factor in the model:

$$\text{Logit}[\text{P}(\text{Contraband found})] = \beta_0 + \beta_1 \text{Race}.$$

We also break down contraband found into three categories: Drugs, Weapons, and Others. We also fit a logistic regression model for each of these categories with `Race` as the sole independent variable.

Logistic Regression for Contraband found

Logistic model for contraband found vs. race

term	estimate	std.error	statistic	p.value
(Intercept)	-1.988	0.222	-8.944	0.000
subject_raceblack	0.899	0.225	3.999	0.000
subject_racehispanic	0.941	0.224	4.202	0.000
subject_racewhite	0.826	0.224	3.686	0.000
subject_raceother	0.796	0.355	2.242	0.025
subject_raceunknown	-0.209	0.416	-0.502	0.616

Logistic regression model for race vs. each category in contraband

Other Results			
Contraband found	Drugs	Weapons	Others
(Intercept)	$-5.24^{***}(1.00)$	$-3.61^{***}(0.45)$	$-2.38^{***}(0.26)$
Black	$1.10(1.01)$	$0.32(0.46)$	$0.99^{***}(0.26)$
Hispanic	$1.21(1.01)$	$0.36(0.46)$	$1.04^{***}(0.26)$
White	$0.73(1.01)$	$0.98^{*}(0.46)$	$0.72^{**}(0.26)$
Other	$0.97(1.42)$	$0.47(0.74)$	$0.87^{*}(0.40)$
Unknown	$-11.33(280.85)$	$0.04(0.85)$	$-0.23(0.53)$

Logistic Regression

- Black and Hispanic drivers are more likely to be frisked than white drivers
  - The estimated odd of being frisked for the black is 2.22 times the estimated odd for the white. This odd ratio for Hispanic people is 1.8.
  - Asian people is the least likely to be frisked.
- Contraband items are more likely to be found from Hispanic and black drivers
  - White people is more likely to be found with weapons
  - Black and Hispanic people are more likely to be found with contraband items that are neither drugs or weapons

Bayesian Modeling

Investigating the Hit Rate

The "hit rate," defined here as the proportion of times an officer finds contraband given that a frisk has been performed, is a widely-used measure for assessing potentially-discriminatory policing. The hit rate can be thought of as a proxy for "evidence" when an officer decides whether to conduct a search or a frisk; a lower hit rate for a particular segment of the population can signal that an officer has a lower threshold of evidence when policing that population segment. In the following analysis, we examine the hit rate at the officer level. Because the analysis requires that officers have stopped all races under consideration, we restrict the analysis to only White, Black, and Hispanic subject races and to officers with 18 or more stops, corresponding to roughly the 90th percentile.

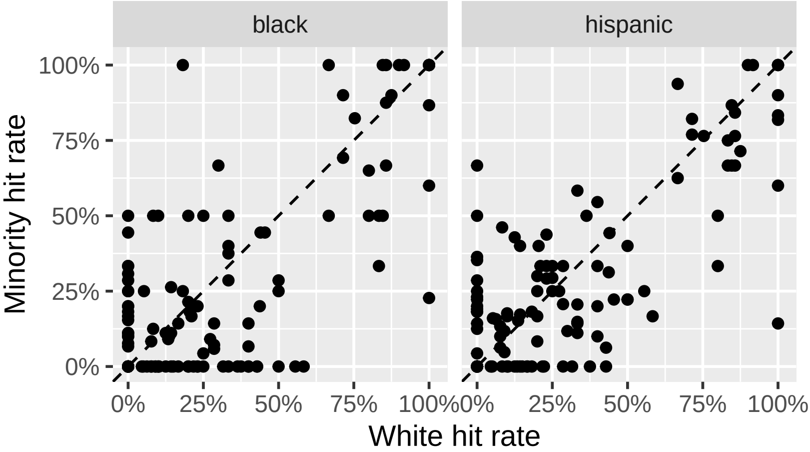


Figure 2: Hit rates for individual officers.

The above plot shows the hit rates for individual officers. An officer with an identical hit rate for white and minority subpopulations would be on the 45-degree line. Visually, it is difficult to determine a systematic trend, although it is clear that particular officers have hit rates that differ

We proceed using a Bayesian hierarchical model. Under this model, we treat individual officers as belonging to a population of players and we seek to model both the hit rates of the officers and the variation of this population. This permits *partial pooling*, by which individual hit rates are biased towards the population average by an amount determined by the estimate of the population. For each officer, we consider three hit rates, one each for white, Black, and Hispanic subpopulations. We accomplish this by fitting separate logistic mixed effects models for each race, each with a weakly informative Normal prior on the log-odds with mean -1.2 and standard deviation 1.

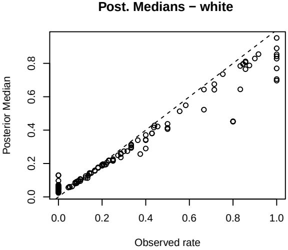
Specifically, let θ_{jr} be the hit rate for for officer j and race r, y_{jr} be the number of hits, and K_{jr} the number of frisks. In the following, because we fit separate models, we assume for example $r = black$ and drop the r subscript. Assuming each officer's searches are independent Bernoulli trials

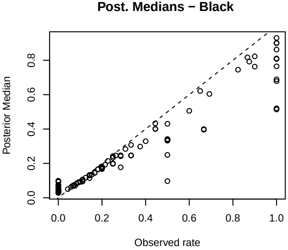
$$p(y_j|\theta_j) = \mathrm{Binomial}(y_j|K_j, \theta_j)$$

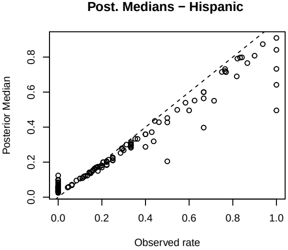
We reparametrize the model in terms of the log-odds, α:

$$\alpha_j = \mathrm{logit}(\theta_j) = \log \frac{\theta_j}{}$$

The the following, the effects of partial pooling are evident: the posterior medians are baised towards the population average. Practically, this means that observed hit rates equal to zero have posterior medians that are small but positive, and perfect (or near-perfect) observed hit rates have somewhat smaller posterior medians.







Because part of this project is to "operationalize" fairness, we devised a measure by which the above posteriors can be converted into a rough "fairness score." Because an officer that uses the same evidence threshold when deciding whether to frisk a subject regardless of race should have roughly equal hit rates for all three subpopulations, we reason that such an officer should have posterior medians that are close to each other for the three subpopulations. So, one can calculate a simple sum of squares statistic for each officer. Specifically, letting m_{jr} be the posterior median for officer j and race r, the sum of squares statistic S_j is

$$S_j = \sum_r (m_{jr} - \bar{m}_j)^2$$

where $\bar{m}_j$ is the average of the three medians. Of course, this measure disregards all other information that could be gleaned from the posterior; an alternative might calculate the overlap between the posterior densities. However, we think this measure is relatively easy to understand and implement.

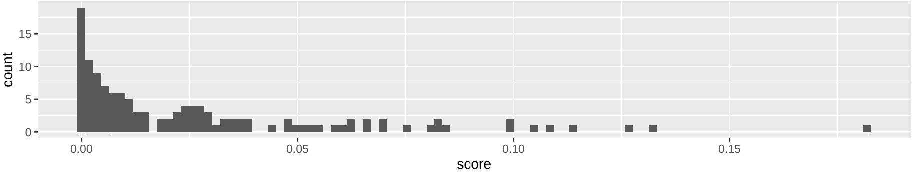


Figure 3: Fairness scores for the officers under consideration. Lower scores indicate hit rates are more similar.

Table 4: Highest 5 scores.

ID	obs.w	obs.b	obs.h	post.w	post.b	post.h	count.w
1504c3bc16	1.000	0.227	0.143	0.697	0.215	0.142	2
3392a495a3	0.400	0.000	0.545	0.371	0.046	0.499	10
50f70c6ecb	0.583	0.000	0.167	0.549	0.064	0.162	12
bab7c2acaf	0.833	0.333	0.750	0.644	0.247	0.715	7
dd9c1003d5	0.556	0.000	0.250	0.513	0.043	0.210	9

Conclusion

- Three Tests (benchmark test, outcome test and veil of darkness test):
  - Evaluate the fairness of traffic stops
  - Confirm racial disparity in policing exists and is present in different scales
- Frequentist Modeling:
  - Explore the causal confounding issues through logistic regression
  - Conclude black and Hispanic people are more likely to be frisked and found with contraband items that are neither drugs or weapons
- Bayesian Modeling:
  - Investigate the hit rate via Bayesian hierarchical modeling
  - Obtain posteriors for the hit rate for each officer in a subset of the data
  - Devised a “fairness score” from the posteriors medians, a tool we believe could be used to identify officers with racially disparate patterns of traffic stops

Thank you for your attention!

UPSTAT

Upstate New York Statistics Conference

FunkyStats: David Skrill, Qiuyi Wu, Cuong Pham (from left to right)

References

Grogger, Jeffrey, and Greg Ridgeway. 2006. *Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness*. Santa Monica, CA: American Statistical Association.

"Hierarchical Partial Pooling for Repeated Binary Trials." n.d. https://mc-stan.org/users/documentation/case-studies/pool-binary-trials.html.

Pierson, Emma, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, et al. 2020. "A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States." *Nature Human Behaviour* 4 (7): 736–45. https://doi.org/10.1038/s41562-020-0858-1.

Simoiu, Camelia, Sam Corbett-Davies, and Sharad Goel. 2017. "THE PROBLEM OF INFRA-MARGINALITY IN OUTCOME TESTS FOR DISCRIMINATION." *The Annals of Applied Statistics* 11 (3): 1193–1216. http://www.jstor.org/stable/26362224.